

On the metric distortion of nearest-neighbour graphs on random point sets

Amitabha Bagchi*

Sohit Bansal*

July 18, 2008

Abstract

We study the graph constructed on a Poisson point process in d dimensions by connecting each point to the k points nearest to it. This graph a.s. has an infinite cluster if $k > k_c(d)$ where $k_c(d)$, known as the critical value, depends only on the dimension d . This paper presents an improved upper bound of 188 on the value of $k_c(2)$. We also show that if $k \geq 188$ the infinite cluster of $\text{NN}(2, k)$ has an infinite subset of points with the property that the distance along the edges of the graphs between these points is at most a constant multiplicative factor larger than their Euclidean distance. Finally we discuss in detail the relevance of our results to the study of multi-hop wireless sensor networks.

1 Introduction

The k -nearest neighbour graph of a point set S in a metric space is constructed according to the following natural definition: For each point $x \in S$ establish an edge from x to the k points of $S \setminus \{x\}$ nearest to it. Such graphs have applications in numerous areas: classification problems of all flavours, topology control in wireless networks [6, 22], data compression [17, 1] and dimensionality reduction [19] and multi-agent systems [10].

We focus on k -nearest neighbor graphs on random point sets in \mathbb{R}^d assuming that the distance is the Euclidean distance. Further we restrict ourselves to the case where the edges established are undirected. Clearly it is not necessary that this graph be connected for arbitrary k and S or even that it have a large connected component. However, Häggström and Meester [13] have shown that if the set S is generated by a Poisson point process then there is a finite value $k_c(d)$ depending only on the dimension such that if $k > k_c(d)$, the k -nearest neighbor graph has an connected component which is infinite. In this paper we study this setting further. Following the notation in [13] we will denote this model in d dimensions, parametrized by k as $\text{NN}(d, k)$.

In this paper we show that for $\text{NN}(2, k)$ that if $k \geq 188$ the infinite cluster¹ has an infinite subset of points with the property the metric distortion between them is bounded by a constant i.e. if there is a pair of points in this infinite subset the shortest distance between them achieved along a path in the graph is at most a constant multiplicative factor larger than the Euclidean distance between them. In the process of proving the latter result we improve the best known bound for $k_c(2)$ to 188 from 213 (due to Teng and Yao [20]). Our proof technique generalizes easily to $\text{NN}(d, k)$ for $d \geq 3$.

*Dept of Computer Science and Engg, Indian Institute of Technology, Hauz Khas, New Delhi 110016. {bagchi,cs503022}@cse.iitd.ernet.in. Note: This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

¹We will use the terms component and cluster interchangeably.

Organization. The rest of this section is devoted to surveying related work and introducing the terms and notation we will use. A new bound on $k_c(2)$ and the result on the metric distortion within the infinite cluster is presented in Section 2. We will talk about the applicability of our results to wireless multi-hop sensor networks in Section 3, concluding with a discussion of some simulation results and conjectures arising from them in Section 4.

1.1 Related work

The study of random graphs obtained by applying connection rules on stationary point processes is known as continuum percolation. Meester and Roy’s monograph on the subject provides an excellent view of the deep theory that has been developed around this general setting [16]. The $NN(d, k)$ model was introduced by Häggström and Meester [13]. They showed that there was a finite critical value, $k_c(d)$ for all $d \geq 2$ such that an infinite cluster exists in this model. They proved that the infinite cluster was unique and that there was a value d_0 such that $k_c(d) = 2$ for all $d > d_0$. Teng and Yao gave an upper bound of 213 for $k_c(d)$ [20].

k -nearest neighbor graphs on random point sets contained inside a finite region have been extensively studied. The major concern, different from ours, has been to ensure that *all* the points within the region are connected within the same cluster. Ballister, Bollobás, Sarkar and Walters [5] showed that the smallest value of k that will ensure connectivity lies between $0.3043 \log n$ and $0.5139 \log n$, improving earlier results of Xue and Kumar [22]. Ballister et. al. also studied the problem of covering the region with the discs containing the k -nearest neighbours of the points. We refer the reader to [5] for an interesting discussion relating this setting to earlier work by Penrose and others.

Eppstein, Paterson and Yao [9] studied k -nearest neighbour graphs on random point sets in two dimensions in some detail and proved interesting bounds showing that the number of points in a component of depth D was polynomial in D when k was 1 and exponential in D when it was 2 or greater. Their primary interest was in obtaining low dilation embeddings of nearest-neighbor graphs.

Algorithms for searching for nearest neighbors (see e.g. [8, 21]) and constructing nearest neighbor graphs efficiently have also received a lot of attention (see e.g. [18]). However these are not directly related so we do not survey this literature in detail.

1.2 Definitions and Notation

Poisson point processes. Our random point sets are generated by homogenous Poisson point processes of intensity λ in \mathbb{R}^d where $d \geq 1$. Under this model the number of points in a region is a random variable that depends only on its d -dimensional volume i.e. the number of points in a bounded, measurable set A is Poisson distributed with mean $\lambda V(A)$ where $V(A)$ is the d -dimensional volume of A . Further, the random variables associated with the number of points in disjoint sets are independent.

Site percolation. Consider an infinite graph defined on the vertex set \mathbb{Z}^d with edges between points x and y such that $\|x - y\|_1 = 1$. Site percolation is a probabilistic process on this graph. Each point of \mathbb{Z}^d is taken to be *open* with probability p and *closed* with probability $1 - p$. The product of all the measures for individual points forms a measure for the space of possible configurations. An edge between two open vertices is considered open. All other edges are considered closed. A component in which open vertices are connected through paths of open edges is known as an open cluster. It is known that there is a value p_c such that for all $p > p_c$ the graph obtained has an infinite open cluster. This value is known as the critical probability. When $p > p_c$ then each point of \mathbb{Z}^d has some non-zero probability of being part of an infinite cluster. The reader is referred to [12] for a full treatment of percolation and to [7] for a recent update on some new directions in this area.

2 An infinite subset of C_∞ has constant metric distortion

The graph distance between pairs of points in a k -nearest neighbor graph is clearly at least the Euclidean distance between them. The question arises if the distance is arbitrarily larger than the Euclidean. Clearly, for points in different clusters the distance this question makes no sense. We also ignore for now the question of what happens inside finite clusters, focussing for now on the infinite cluster in the supercritical phase of $NN(d, k)$. We conjecture that it is possible to show that the distance between any pair of points in the infinite cluster is only a small factor larger than the Euclidean distance between them. In this paper we prove a weaker result: the infinite cluster contains an infinite subset of points whose pairwise distances are not distorted by more than a constant factor. In order to do this we first present a construction that allows us to couple $NN(2, k)$ with a site percolation process in \mathbb{Z}^2 . This construction also improves the best known upper bound for $k_c(2)$. Then we show how to use the algorithm of Angel et. al. [2] for routing on a percolated mesh to find a short path between a pair of vertices in $NN(d, k)$.

2.1 Coupling $NN(2, k)$ to site percolation in \mathbb{Z}^2

Like Häggström and Meester's proof for the existence of a critical value [13] and Teng and Yao's proof for the weaker of their two upper bounds on $k_c(2)$ [20], we proceed by constructing a coupling with a site percolation process on \mathbb{Z}^2 . However, our construction gives a better upper bound than Teng and Yao's improvement of their own result (also in [20]) to $k_c(2) \geq 213$ which uses a coupling to a mixed percolation process. We are able to improve this result to show $k_c(2) \geq 188$. Note that both papers, the one by Häggström and Meester and the one by Teng and Yao, reported that simulations seemed to indicate that the value of $k_c(2)$ appears to be around 3. Our simulations backed up this finding. Let us now proceed to a formal statement of the main theorem of this section and its proof.

Theorem 2.1 *For the k -nearest neighbour model in a Poisson point process setting*

$$k_c(2) \leq 188.$$

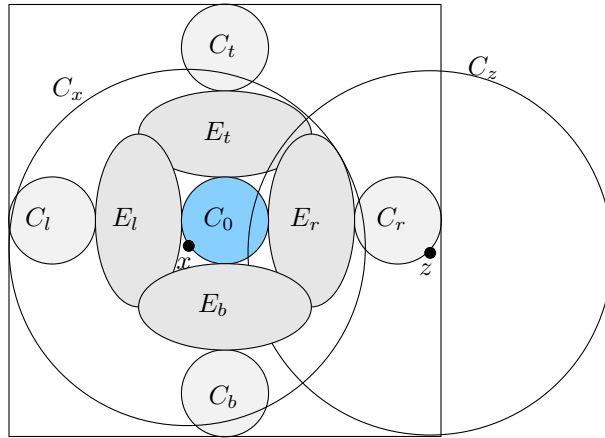


Figure 1: A tile t and its 9 relevant regions. Note that the region E_r lies wholly within all discs of the form C_x and C_z centred at points on the boundary of the discs C_0 and C_r .

Proof. In order to prove the theorem we couple a site percolation process on \mathbb{Z}^2 with the k -nearest neighbour graph as follows. We divide \mathbb{R}^2 into square tiles of side $10a$ where a is a parameter whose

value will be fixed later. We create a bijection, ϕ , between these tiles in \mathbb{R}^2 and points in \mathbb{Z}^2 such that neighbouring tiles in \mathbb{R}^2 correspond to neighbouring points in \mathbb{Z}^2 . We couple the processes by saying that a given point x in \mathbb{Z}^2 is open only if the tile $t = \phi^{-1}(x)$ a certain event A_t occurs. We now define this event A_t .

Let us look at a tile centred at $(0,0)$ with bottom left corner $(-5a, -5a)$ and top right corner $(5a, 5a)$. For convenience we will refer to the tiles surrounding the tile t as, counterclockwise starting from the right t_r, t_t, t_l and t_b . We consider five circles of radius a : C_0 centred at $(0,0)$, C_l centred at $(-4a, 0)$, C_r centred at $(4a, 0)$, C_t centred at $(0, 4a)$ and C_b centred at $(0, -4a)$. There are four other regions which are named E_l, E_r, E_t and E_b in the figure. E_r is defined as follows. Consider the largest circle centred at any point in C_0 or C_r that lies wholly within the two tiles t and t_r . Two such circles, C_x and C_z , are depicted in Figure 1. E_r is the locus of the points contained in all such circles. The regions E_l, E_t and E_b are defined similarly by C_0 along with C_l, C_t and C_b respectively and the tiles t_l, t_t and t_b respectively.

Now, for a tile t , the event A_t is said to occur if

1. the number of points inside t is at most $k/2$ and
2. the nine regions $C_0, C_r, C_t, C_l, C_b, E_r, E_t, E_l$ and E_b contain at least one point each.

If A_t occurs we call the point contained in C_0 the *representative point* of the tile t , denoted $\text{rep}(t)$. In order to relate the process on \mathbb{Z}^2 defined via these events A_t to the $\text{NN}(d, k)$ model, we claim that the existence of an edge in \mathbb{Z}^2 implies the existence of a path from the representative points of the two tiles corresponding to the two end points of the edge. We state this formally, including an observation about the metric distortion of the length of the path between the two representative points.

Claim 2.2 *If an edge exists in the percolated mesh \mathbb{Z}^2 between two points x and y then*

1. *There is a path between the representative points $\text{rep}(\phi^{-1}(x))$ and $\text{rep}(\phi^{-1}(y))$ of the tiles corresponding to x and y in $\text{NN}(2, k)$ and*
2. *there is a constant c_{tiles} such that*

$$d_k(\text{rep}(\phi^{-1}(x)), \text{rep}(\phi^{-1}(y))) \leq c \cdot d(\text{rep}(\phi^{-1}(x)), \text{rep}(\phi^{-1}(y))).$$

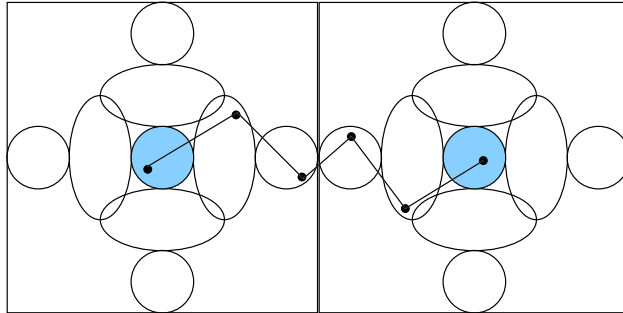


Figure 2: A path between two representative points of tiles for both of which the event A_t has occurred.

Proof of Claim 2.2: The proof of the claim is depicted in Figure 2. Clearly any circle drawn from $\text{rep}(t)$ that stays within t contains all of E_r in it by the definition of E_r . Since there are at most $k/2$

points in every tile for which A_t has occurred, hence there is an edge from $\text{rep}(t)$ to the point guaranteed to be contained in E_r , let's call it x_r , by the definition of A_t . We do not make any claims on where the edges established by x_r to its neighbours lie, observing only that any point that lies in C_r must have an edge to x_r , again by the definition of E_r . However, any disc centred at a point in C_r that remains within t and t_r must contain the left disc of its neighboring tile. Hence, if A_t and A_{t_r} occur then a path from $\text{rep}(t)$ to $\text{rep}(t_r)$ occurs. The second part of the claim is obviously true. The constant can easily be calculated using calculus. \square

From Claim 2.2, it is easy to deduce that if an infinite component exists in the site percolation setting, then an infinite component exists in $\text{NN}(2, k)$. Hence we need to determine for what settings of our parameters a and, more importantly, k , the site percolation process is supercritical. The critical probability for site percolation is 0.59 (see e.g. [15]). Numerical calculations showed that the smallest value of k for which the probability of A_t exceeds this value is 188, and the value of a for which this happens is 0.893. \square

2.2 A subset with constant metric distortion

We now show that there is a set of points in C_∞ and constant α such that for each pair of points x, y in this set

$$D_k(x, y) \leq \alpha \cdot D(x, y).$$

We will prove the following theorem

Theorem 2.3 *For $\text{NN}(2, k)$ where $k > 188$, there is a set of points $S \subseteq C_\infty$ such that $|S| = \infty$ with the following property: Let $x, y \in S$ be two points with Euclidean distance $D(x, y)$ between them whose k -NN distance is $D_k(x, y)$. For some α, c depending only on k*

$$P(D_k(x, y) > \alpha \cdot D(x, y)) < e^{-c \cdot D(x, y)}.$$

Proof. We identify S to be the set of representative points lying in the infinite cluster of $\text{NN}(2, k)$ of the construction described in the proof of Theorem 2.1 as the subset that we will claim has this property. We use the coupling with site percolation in \mathbb{Z}^2 introduced in that proof to help us find short paths between pairs of points in S .

Let us consider any two tiles t_1 and t_2 whose representative points $\text{rep}(t_1)$ and $\text{rep}(t_2)$ lie in C_∞ . We denote distance between two points a, b in \mathbb{Z}^2 is denoted $D_{\text{latt}}(a, b)$. First we relate the distance in the (unpercolated) lattice to the euclidean distance between these two points by observing a simple fact.

Fact 2.4 *Given that c_{tiles} is the constant defined in Claim 2.2 then for two tiles t_1, t_2*

$$D_{\text{latt}}(\phi(\text{rep}(t_1)), \phi(\text{rep}(t_2))) \leq \sqrt{2} \cdot \frac{D(\text{rep}(t_1), \text{rep}(t_2))}{c_{\text{tiles}}}.$$

When the lattice undergoes percolation, the simple open path from $\phi(\text{rep}(t_1))$ to $\phi(\text{rep}(t_2))$ may be broken at several points. Antal and Pisztora studied this setting and proved a powerful theorem which helps us here [3, Theorems 1.1 and 1.2]. We use it as a lemma here, adopting the restatement of Angel et. al. [2, Lemma 8].

Lemma 2.5 [3, 2] *For any $p > p_c$ and any x, y connected through an open path in a cube M^d of the infinite lattice, let $D_{\text{latt}}^p(x, y)$ be the distance between the two points in the percolated lattice. For some $\rho, c_2 > 0$ depending only on the dimension and p and for any $a > \rho \cdot D_{\text{latt}}(x, y)$*

$$\text{pr}(D_{\text{latt}}^p(x, y) > a) < e^{-c_2 a}.$$

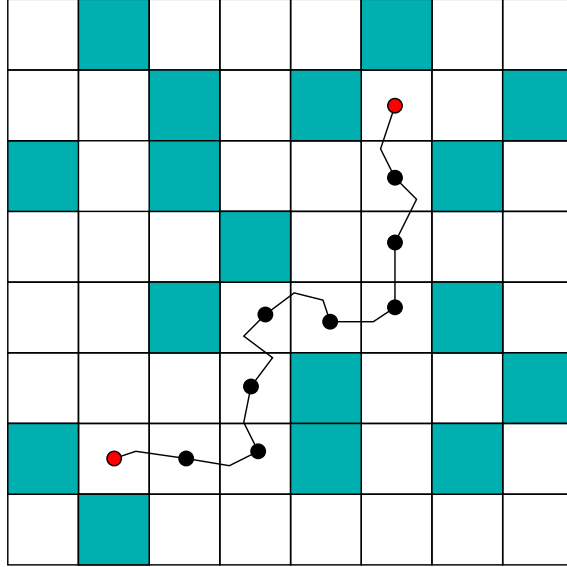


Figure 3: The path between two representative points mimics the path in \mathbb{Z}^2 .

To find a path between $\text{rep}(t_1)$ and $\text{rep}(t_2)$ we simply take the path in the percolated lattice between $\phi(t_1)$ and $\phi(t_2)$ and mimic it \mathbb{R}^2 as depicted in Figure 3 and the result follows by combining Fact 2.4 and Lemma 2.5. \square

We note here that our claim that the constant in the statement of Theorem 2.3 depends only on the value of k follows from the fact that the constants in Lemma 2.5 depend only on p , since in our construction the size of a tile and the probability of A_t occurring for a tile changes when we change k . We also note that Antal and Pisztora [3] prove their theorem for bond percolation but note that their methods can easily be extended to site percolation.

It is possible to extend Theorem 2.3 easily for $d > 2$. The constants change and their dependence on d has to be handled carefully but the proof remains basically the same.

3 Applications to multi-hop wireless sensor networks.

Multi-hop sensor networks, where nodes act not only to sense but also to relay information, have proven advantages in terms of energy efficiency over single hop sensor networks [14] and are useful necessary tasks like time synchronization [11]. And for collaborative tasks like target tracking [23] sensor-to-sensor communication is essential. But the total connectivity sought to be achieved in [22, 5] between all the points of a point process is not necessary for these networks. It may be the correct model for general ad hoc wireless networks where all nodes need to be connected, but for a sensor network we argue the presence of large connected component is enough.

Sensor networks seek to achieve coverage of a target area. When the locations of sensors are modelled by point processes achieving most coverage measures (whether it is single point coverage or k -coverage or barrier coverage) has found that there is a critical density of the point process above which the particular measure is satisfactory. For example [4] estimates the critical density required for barrier coverage in strip-like regions, a notion of coverage where an object must be sensed if it tries to

n	k	avg.	max. value	percentage
500	3	1.727	15.180	96.96%
500	4	1.364	7.543	97.96%
500	5	1.204	5.874	99.38%
1000	3	1.660	22.64	97.08%
1000	4	1.333	8.39	98.92%
1000	5	1.172	4.385	99.82%
1500	4	1.322	7.858	99.12%
2000	4	1.285	9.512	99.4%

Table 1: Metric distortion in $NN(2, k)$. The last column shows the percentage of pairs distorted by a factor of 2 or less.

cross a particular region.²

Our results show that it is possible to find an infinite component with which has an infinite subset of nodes whose graph distance is a constant times their euclidean distance. Our construction for the proofs of Theorems 2.1 and 2.3, taken along with the fact that for any point in \mathbb{Z}^2 there is a non-zero probability of being part of the infinite component in the supercritical phase imply the following theorem

Theorem 3.1 *For any λ , there is a λ' such that $NN(2, k)$ built on a point process of density λ' with $k > 188$, has an infinite component with the property that an infinite subset of points with density at least λ has the property that that graph distance between them is at most a constant times the Euclidean distance between them. Moreover there is a constant c such that $\lambda' < c\lambda$.*

Clearly the existence of such a subset can fulfil the sensing requirement while not compromising on the sensor-to-sensor data transfer capability. The value 188 seems prohibitive for most practical purposes. But it is our hope that this upper bound will be improved down to a reasonable value closer to the 2 conjectured by Häggström and Meester [13] and Teng and Yao [20] and that it will be possible to prove Theorem 2.3 for this improved bound as well. We omit the proof of this theorem here since it does not add any new insight over the proofs already seen in this paper.

4 Conclusion

We conclude by presenting some conjectures about the relationship of the metric distortion in $NN(2, k)$ to the parameter k . These conjectures come from simulations we ran.

The experiments had to be carried out on a finite box in \mathbb{R}^2 , but to negate boundary effects we simulated a point process in a large box and looked at the largest component formed within a smaller box contained well within this finite box. We placed a number of points randomly within the larger area (thereby achieving a target density). In Table 4 the first column has the number of points placed. The table shows the average distortion for different values of k , maximum value of the distortion and the percentage of points having distortion less than two times the average. This table also indicates that there the distortion is independent of the number of points under consideration but depends on the value of k .

To show relationship between k and average distortion we plotted average ratio with k^2 for a range of value of k from 3 to 13 for two random point sets. Figures 4 and 4 show plots for two such sets

²See [14, Chap 13.2] for a succinct summary of the issues involved in coverage.

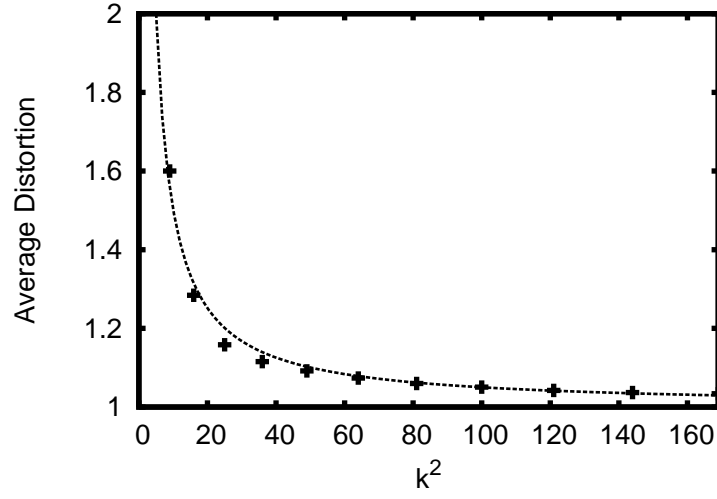


Figure 4: Average metric distortion on the y -axis and k^2 on the x axis. The curve plotted is $1 + 4.62/k^2$.

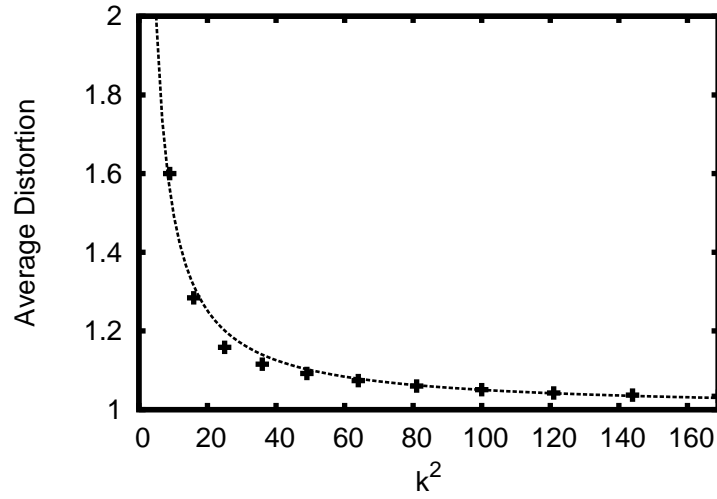


Figure 5: Average metric distortion on the y -axis and k^2 on the x axis. The curve plotted is $1 + 5.03/k^2$.

along with a function $f(k) = 1 + a/k^2$ where a is determined by least square fitting functions. These findings lead us to conjecture that:

Conjecture 4.1 *For the $NN(2, k)$ model at a value $k > k_c(2)$*

1. *The metric distortion of the points of C_∞ is at most 2 with probability tending to 1 and*
2. *there is a constant such that the expected metric distortion of the points of C_∞ is of the form $1 + \frac{a}{k^2}$.*

References

- [1] M. Adler and M. Mitzenmacher. Towards compressing web graphs. In *Proc. of the IEEE Data Compression Conference (DCC)*, pages 203–212, 2001.
- [2] O. Angel, I. Benjamini, E. Ofek, and U. Wieder. Routing complexity of faulty networks. In *Proc. of 24th Annu. ACM Symp. on Principles of Distributed Computing (PODC 2005)*, pages 209–217, 2005.
- [3] P. Antal and A. Pisztora. On the chemical distance for supercritical Bernuolli percolation. *Ann. Probab.*, 24(2):1036–1048, 1996.
- [4] P. Ballister, B. Bollobás, A. Sarkar, and S. Kumar. Reliable density estimates for coverage and connectivity in thin strips of finite length. In *Proc. 13th Annual ACM intl. conf. on Mobile computing and networking (MOBICOM 2007)*, pages 75–86, 2007.
- [5] P. Ballister, B. Bollobás, A. Sarkar, and M. Walters. Connectivity of random k -nearest-neighbour graphs. *Adv. Appl. Prob. (SGSA)*, 37:1–24, 2005.
- [6] D. M. Blough, M. Leoncini, G. Resta, and P. Santi. The k -neigh protocol for symmetric topology control in ad hoc networks. In *Proc. 4th Intl. Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2003)*, pages 141–152, 2003.
- [7] B. Bollobás and O. Riordan. *Percolation*. Cambridge University Press, 2006.
- [8] K. Clarkson. Nearest-neighbor searching and metric space dimensions. In *Methods for Learning and Vision: Theory and Practice*. MIT Press, 2005.
- [9] D. Eppstein, M. Paterson, and F. F. Yao. On nearest-neighbor graphs. *Discrete Comput. Geom.*, 17(3):263–282, 1997.
- [10] A. Goebels. Studies on neighbourhood graphs for communication in multi agent systems. In *Proc. (part II) Advances in Natural Computation, 2nd Intl. Conf. (ICNC 2006)*, pages 456–465, 2006.
- [11] J. V. Greunen and J. M. Rabaey. Lightweight time synchronization in sensor networks. In *Proc. 2nd ACM Intl. Conf. on Wireless Sensor Networks and Applications (WSNA 2003)*, pages 11–19, 2003.
- [12] G. Grimmett. *Percolation*, volume 321 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2nd edition, 1999.
- [13] O. Häggström and R. Meester. Nearest neighbor and hard sphere models in continuum percolation. *Random Struct. Algor.*, 9(3):295–315, 1996.

- [14] H. Karl and A. Willig. *Protocols and Architectures for wireless sensor networks*. John Wiley and Sons, 2005.
- [15] M. J. Lee. Complementary algorithms for graphs and percolation. arXiv:0708.0600v1, 2007.
- [16] R. Meester and R. Roy. *Continuum Percolation*. Number 119 in Cambridge Tracts in Mathematics. Cambridge University Press, 1996.
- [17] Z. Ouyang, N. Memon, T. Suel, and D. Trendafilov. Cluster-based delta compression of a collection of files. In *Proc. 3rd Intl. Conf. on Web Information Systems Engineering*, pages 257–268, 2002.
- [18] R. Paredes, E. Chávez, K. Figueroa, and G. Navarro. Practical construction of k -nearest neighbor graphs in metric spaces. In *Proc. 5th Intl. Workshop on Experimental Algorithms (WEA 2006)*, pages 85–97, 2006.
- [19] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.
- [20] S.-H. Teng and F. F. Yao. k -nearest-neighbor clustering and percolation theory. *Algorithmica*, 49:192–211, 2007.
- [21] P. M. Vaidya. An $o(n \log n)$ algorithm for the All-Nearest-Neighbors problem. *Discrete Comput. Geom.*, 4:101–115, 1989.
- [22] F. Xue and P. R. Kumar. The number of neighbours needed for the connectivity of wireless networks. *Wireless Networks*, 10:169–181, 2004.
- [23] F. Zhao, J. Liu, L. Guibas, and J. Reich. Collaborative signal and information processing: An information directed approach. *Proc. IEEE*, 32(1):61–72, 2003.